METHOD FOR EPIGENETIC FEATURE SELECTION

[0001] This application is a continuation-in-part of application number 10/106,269, filed March 26, 2002, which claims priority to provisional application number 60/278,333, filed on March, 26, 2001. Both the 10/106,269 and 60/278,333 applications are hereby incorporated by reference herein. All references cited in the present application are hereby incorporated by reference herein.

[0002] The present invention is related to methods and computer program products for biological data analysis. Specifically, the present invention relates to methods and computer program products for the analysis of large scale DNA methylation data.

BACKGROUND

[0003] The levels of observation that have been well studied by the methodological developments of recent years in molecular biology, are the genes themselves, the translation of these genes into RNA, and the resulting proteins. Many biological functions, disease states and related conditions are characterized by differences in the expression levels of various genes. These differences may occur through changes in the copy number of the genomic DNA, through changes in levels of transcription of the genes, or through changes in protein synthesis.

[0004] Recently, massive parallel gene expression monitoring methods have been developed to monitor the expression of a large number of genes using mRNA based nucleic acid microarray technology (see, *e.g.,* Lockhart, D.J. *et al.*, Expression monitoring by hybridization to high density Oligonucleotid arrays, *Nature Biotechnology* 14:1675-1680, 1996; Lockhart, D.J. *et al.*, Genomics, gene expression and DNA arrays, *Nature* 405:827-836, 2000). This technology allows to look at thousands of genes simultaneously, see how they are expressed as proteins and gain insight into cellular processes.

[0005] However, large scale analysis using mRNA based microarrays are primarily impeded by the instability of mRNA (Emmert-Buck, T. *et al.*, *Am J Pathol.* 156, 1109, 2000). Also expression changes of only a minimum of a factor 2 can be routinely and reliably detected

(Lipshutz, R. J. *et al.*, High density synthetic oligonucleotide arrays, *Nature Genetics* 21, 20, 1999; Selinger, D. W. *et al.*, RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array, *Nature Biotechnology* 18, 1262, 2000). Furthermore, sample preparation is complicated by the fact that expression changes occur within minutes following certain triggers.

[0006] An alternative approach is to look at DNA methylation. 5-methylcytosine is the most frequent covalent base modification in the DNA of eukaryotic cells. It plays a role, for example, in the regulation of the transcription, in genetic imprinting, and in tumorigenesis. For example, aberrant DNA methylation within CpG islands is common in human malignancies leading to abrogation or overexpression of a broad spectrum of genes (Jones, P.A., DNA methylation errors and cancer, *Cancer Res.* 65:2463-2467, 1996). Abnormal methylation has also been shown to occur in CpG rich regulatory elements in intronic and coding parts of genes for certain tumors (Chan, M.F., *et al.*, Relationship between transcription and DNA methylation, *Curr. Top. Microbiol. Immunol.* 249:75-86,2000). Using restriction landmark genomic scanning, Costello and coworkers were able to show that methylation patterns are tumour-type specific (Costello, J. F. *et al.*, Aberrant CpG-island methylation has non-random and tumor-type-specific patterns, *Nature Genetics* 24:132-138, 2000). Highly characteristic DNA methylation patterns could also be shown for breast cancer cell lines (Huang, T. H.-M. *et al.*, *Hum. Mol. Genet.* 8:459-470, 1999).

[0007] Therefore, the identification of 5-methylcytosine as a component of genetic information is of considerable interest. However, 5-methylcytosine positions cannot be identified by sequencing since 5-methylcytosine has the same base pairing behavior as cytosine. Moreover, the epigenetic information carried by 5-methylcytosine is completely lost during PCR amplification.

[0008] The state of the art method for large scale methylation analysis (PCT Publication No. WO99/28498) is based upon the specific reaction of bisulfite with cytosine which, upon subsequent alkaline hydrolysis, is converted to uracil which corresponds to thymidine in its base pairing behavior. However, 5-methylcytosine remains unmodified under these conditions. Consequently, the original DNA is converted in such a manner that methylcytosine, which originally could not be distinguished from cytosine by its hybridization behavior, can now be detected as the only remaining cytosine using "normal" molecular biological techniques, for

example, by amplification and hybridization to oligonucleotide microarrays or sequencing.

[0009] Like mRNA based massive parallel gene expression monitoring experiments, large scale methylation analysis experiments generate unprecedented amounts of information. A single hybridization experiment can produce quantitative results for thousands of CpG positions. Therefore, there is a great need in the art for methods and computer program products to organize, access and analyze the vast amount of information collected using large scale methylation analysis methods.

[0010] One approach is to use unsupervised or supervised machine learning methods to analyze large scale methylation data. However, in large scale methylation analysis the extreme high dimensionality of the data compared to the usually small number of available samples is a severe problem for all classification methods. Therefore, for good performance of the machine learning methods a reduction of the data dimensionality is necessary. This problem is solved by the present invention. The invention provides methods and computer program products for the selection of epigenetic features, as for example the methylation status of CpG positions. Only the corresponding data to these epigenetic features is then subject to machine learning analysis thereby crucially improving the performance of the machine learning analysis.

SUMMARY OF THE INVENTION

[0011] The present invention provides methods and computer program products for selecting epigenetic features. The methods and computer program products are particularly useful in large scale nucleic acid methylation analysis.

[0012] In one aspect of the invention methods are provided for selecting epigenetic features comprising the following steps:

[0013] In the first step, biological samples containing genomic DNA are collected and stored. The biological samples may comprise cells, cellular components which contain DNA or free DNA. Such sources of DNA may include cell lines, biopsies, blood, sputum, stool, urine, cerebral-spinal fluid, tissue embedded in paraffin such as tissue from eyes, intestine, kidney, brain, heart, prostate, lung, breast or liver, histologic object slides, and all possible combina-

tions thereof.

[0014] Next, available phenotypic information about said biological samples is collected and stored, thereby defining a phenotypic data set for the biological samples. The phenotypic information may comprise, for example, kind of tissue, drug resistance, toxicology, organ type, age, life style, disease history, signaling chains, protein synthesis, behavior, drug abuse, patient history, cellular parameters, treatment history and gene expression.

[0015] Next, at least one phenotypic parameter of interest is defined. These defined phenotypic parameters of interest are used to divide the biological samples in at least two disjunct phenotypic classes of interest.

[0016] An initial set of epigenetic features of interest is defined. Epigenetic features of interest are, for example, cytosine methylation statuses at selected CpG positions in DNA. This initial set of epigenetic features of interest may be defined using preliminary knowledge data about their correlation with phenotypic parameters.

[0017] The defined epigenetic features of interest of the biological samples are measured and/or analyzed, thereby generating an epigenetic feature data set.

[0018] Next, those epigenetic features of interest and/or combinations of epigenetic features of interest are selected that are relevant for epigenetically based prediction of the phenotypic classes of interest. An epigenetic feature of interest and/or combination of epigenetic features of interest is preferably considered relevant for epigenetically based class prediction if the accuracy and/or the significance of the epigenetically based prediction of said phenotypic classes of interest is likely to decrease by exclusion of the corresponding epigenetic feature data.

[0019] Finally, a new set of epigenetic features of interest is defined based on the relevant epigenetic features of interest and/or combinations of epigenetic features of interest generated in the preceding step.

[0020] In some embodiments of the invention the steps of measuring and/or analyzing the epigenetic features of interest of the biological samples and of selecting the relevant epigenetic features of interest are iteratively repeated based on the epigenetic features of interest

defined in the preceding iteration.

[0021] In one preferred embodiment, the phenotypic parameters of interest are used to divide the biological samples in two disjunct phenotypic classes of interest. In this embodiment, a machine learning classifier may be used for epigenetically based prediction of the two disjunct phenotypic classes of interest. In another preferred embodiment, the disjunct phenotypic classes of interest are grouped in pairs of classes or pairs of unions of classes and machine learning classifiers may be applied for epigenetically based class prediction to each pair.

[0022] In preferred embodiments the selection of the relevant epigenetic features of interest and/or combinations of epigenetic features of interest is done by a) defining a candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest, b) defining a feature selection criterion, c) ranking the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest according to the defined feature selection criterion and d) selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

[0023] The defined candidate set of epigenetic features of interest may be the set of all subsets of the epigenetic features of interest, preferably the set of all subsets of a given cardinality of said defined epigenetic features of interest, in a preferred embodiment the set of all subsets of cardinality 1.

[0024] In another preferred embodiment the measured and/or analyzed epigenetic feature data set is subject to principal component analysis, the principal components defining a candidate set of linear combinations of the defined epigenetic features of interest.

[0025] In other embodiments dimension reduction techniques preferably multidimensional scaling, isometric feature mapping or cluster analysis are used to define the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. The cluster analysis may be hierarchical clustering or k-means clustering.

[0026] In a preferred embodiment of the method the candidate set of epigenetic features of interest is determined based on a priori biological information such that epigenetic features of interest with common biological properties are grouped together to form the candidate set of epigenetic features. It is preferred that said common biological properties (also referred to

herein as 'biological factors') are a common methylation status, known in the field as 'co-methylation'. Wherein this is not known it may be inferred using any parameters which may be used as reasonable indicators that members of a set of CpG positions have a common methylation status, which may in particular be selected from the group consisting of:

- Proximity to each other; wherein the epigenetic features are close enough that it may be assumed or expected that they have similar or correlated epigenetic status. In particular, when the epigenetic features belong to the same CpG island (defined as a sequence greater than 200 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65 or greater). (Taken from D.T and P.A.J.[PNAS 99(6):3740-5 (2002)]); and

- Associated function: epigenetic features belonging to genes that are known to have similar function and/or co-regulated and/or belong to the same biological pathway and/or have sequence similarity and therefore expected to be regulated by similar transcription factors.

[0027] In preferred embodiments which use machine learning classifiers for the prediction of the phenotypic classes of interest based on the epigenetic feature data set the feature selection criterion may be the training error of the machine learning classifier trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In another preferred embodiment the epigenetic feature selection criterion may be the risk of the machine learning classifier trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In a further preferred embodiment, the epigenetic feature selection criterion may be the bounds on the risk of the machine learning classifier trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

[0028] In preferred embodiments in which the candidate set of epigenetic features of interest comprises single epigenetic features or single combinations of epigenetic features of interest the epigenetic feature selection criterion may be the use of test statistics for computing the significance of difference of the phenotypic classes of interest given the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Preferably the statistical test may be a t-test or a

rank test, for example a Wilcoxon rank test. In a preferred embodiment of the method the statistical test used for combining the epigenetic features is a multivariate statistical test, suitable tests include but are not limited to Hotelling's T^2 test and the likelihood ratio test for logistic regression models. In one preferred embodiment, the epigenetic feature selection criterion may be the computation of the Fisher criterion for the phenotypic classes of interest given the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Furthermore the epigenetic feature selection criterion may be the computation of the weights of a linear discriminant for said phenotypic classes of interest given the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Preferred linear discriminants are the Fisher discriminant or the discriminant of a support vector machine classifier for said phenotypic classes of interest trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In yet another embodiment, the epigenetic feature selection criterion may be subjecting the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest to principal component analysis and calculating the weights of the first principal component. Moreover, the epigenetic feature selection criterion can be chosen to be the mutual information between the phenotypic classes of interest and the classification achieved by an optimally selected threshold on the given epigenetic feature of interest. Still further, the epigenetic feature selection criterion may be the number of correct classifications achieved by an optimally selected threshold on the given epigenetic feature of interest.

[0029] In preferred embodiments in which the epigenetic feature data set is subject to principal component analysis, the principal components defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest, the feature selection criterion can be chosen to be the eigenvalues of the principal components.

[0030] In a preferred embodiment wherein the candidate set of epigenetic features of interest comprises single epigenetic features or single combinations of epigenetic features of interest the epigenetic feature selection criterion may be the average degree of methylation of the given epigenetic feature or set of epigenetic features on a given subset of samples. In one preferred embodiment, the epigenetic feature selection criterion may be the computation of

the average degree of methylation on peripheral blood samples.

[0031] In preferred embodiments in which the candidate set of epigenetic features of interest comprises single combinations of epigenetic features of interest the epigenetic feature selection criterion may be the average pairwise correlation between the single epigenetic features.

[0032] In some preferred embodiments, the epigenetic features of interest and/or combinations of epigenetic features of interest selected may be a defined number of the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest. In other preferred embodiments, all except a defined number of lowest ranking epigenetic features of interest and/or combinations of epigenetic features of interest are selected. In yet other preferred embodiments, the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected or all except the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score lesser than a defined threshold are selected.

[0033] In preferred embodiments, the iterative method of the invention is repeated until a defined number of epigenetic features of interest and/or combinations of epigenetic features of interest are selected or until all epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected.

[0034] In preferred embodiments the optimal number of epigenetic features of interest and/or combinations of epigenetic features of interest and/or the optimal feature selection criterion score threshold is determined by crossvalidation of a machine learning classifier on test subsets of the epigenetic feature data.

[0035] In some embodiments of the invention, the feature data set corresponding to the defined new set of epigenetic features of interest is used to train a machine learning classifier.

[0036] In another aspect of the invention computer program products are provided. An exemplary computer program product comprises: a) computer code that receives as input an epigenetic feature data-set for a plurality of epigenetic features of interest, the epigenetic feature data-set being grouped in disjunct classes of interest; b) computer code that selects those epi-

genetic features of interest and/or combinations of epigenetic features of interest that are relevant for machine learning class prediction based on the epigenetic feature data set; c) computer code that defines a new set of epigenetic features of interest based on the relevant epigenetic features of interest and/or combinations of epigenetic features of interest generated in step (b); d) a computer readable medium that stores the computer code. In a preferred embodiment, the computer code repeats step (b) iteratively based on the new defined set of epigenetic features of interest defined in step (c).

[0037] Preferably, an epigenetic feature of interest and/or combination of epigenetic features of interest are considered relevant for machine learning class prediction if the accuracy and/or the significance of the class prediction is likely to decrease by exclusion of the corresponding epigenetic feature data.

[0038] In one preferred embodiment, the computer code groups the epigenetic feature data set in disjunct pairs of classes and/or pairs of unions of classes of interest before applying the computer code of steps (b) and (c).

[0039] In preferred embodiments the computer code selects the relevant epigenetic features of interest and/or combinations of epigenetic features of interest by a) defining candidate sets of epigenetic features of interest and/or combinations of epigenetic features of interest b) ranking the candidate sets of epigenetic features of interest and/or combinations of epigenetic features of interest according to a feature selection criterion and c) selecting the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest.

[0040] The candidate set of epigenetic features of interest the computer code chooses for ranking may be the set of all subsets of the epigenetic features of interest, preferably the set of all subsets of a given cardinality, particularly the set of all subsets of cardinality 1.

[0041] In another preferred embodiment the computer code subjects the epigenetic feature data set to principal component analysis, the principal components defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

[0042] In other embodiments the computer code applies dimension reduction techniques preferably multidimensional scaling, isometric feature mapping or cluster analysis to define the candidate set of epigenetic features of interest and/or combinations of epigenetic features of

interest. The cluster analysis may be hierarchical clustering or k-means clustering.

[0043] In a preferred embodiment of the method the computer code determines the candidate set of epigenetic features of interest based upon a priori biological information such that epigenetic features of interest with common biological properties are grouped together to form the candidate set of epigenetic features. It is preferred that said common biological properties (also referred to herein as 'biological factors') are a common methylation status, known in the field as 'co-methylation'. Wherein this is not known it may be inferred using any parameters which may be used as reasonable indicators that members of a set of CpG positions have a common methylation status, which may in particular be selected from the group consisting of:

- Proximity to each other; wherein the epigenetic features are close enough that it may be assumed or expected that they have similar or correlated epigenetic status. In particular, when the epigenetic features belong to the same CpG island (defined as a sequence greater than 200 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65 or greater); and

- Associated function: epigenetic features belonging to genes that are known to have similar function and/or co-regulated and/or belong to the same biological pathway and/or have sequence similarity and therefore expected to be regulated by similar transcription factors. In preferred embodiments the feature selection criterion used by the computer code may be the training error of the machine learning classifier algorithm trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In another preferred embodiment the epigenetic feature selection criterion is the risk of the machine learning classifier algorithm trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In a further preferred embodiment, the epigenetic feature selection criterion are the bounds on the risk of the machine learning classifier trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest.

[0044] In preferred embodiments in which the candidate set of epigenetic features of interest defined by the computer code comprises single epigenetic features or single combinations of

epigenetic features of interest the epigenetic feature selection criterion used by the computer code may be the use of test statistics for computing the significance of difference of the classes of interest given the epigenetic feature data corresponding to the chosen candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Preferably the statistical test may be a t-test or a rank test, for example a Wilcoxon rank test. Most preferably the combination of epigenetic features by computer code is carried out by means of a multivariate statistical test, suitable tests include, but are not limited to, a Hotelling's T^2 test or the likelihood ratio test for logistic regression models. In one preferred embodiment, the epigenetic feature selection criterion may be the computation of the Fisher criterion for the classes of interest given the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Furthermore the epigenetic feature selection criterion may be the computation of the weights of a linear discriminant for the classes of interest given the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. Preferred linear discriminants are the Fisher discriminant or the discriminant of a support vector machine classifier for the classes of interest trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. In yet another embodiment, the computer code subjects the epigenetic feature data corresponding to the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest to principal component analysis and calculates the weights of the first principal component as feature selection criterion. Moreover, the epigenetic feature selection criterion can be chosen to be the mutual information between the classes of interest and the classification achieved by an optimally selected threshold on the given epigenetic feature of interest. Still further, the epigenetic feature selection criterion may be the number of correct classifications achieved by an optimally selected threshold on the given epigenetic feature of interest.

[0045] In preferred embodiments in which the computer code subject the epigenetic feature data set to principal component analysis, the principal components defining the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest, the feature selection criterion can be chosen to be the eigenvalues of the principal components.

[0046] In a preferred embodiment wherein the candidate set of epigenetic features of interest

comprises single epigenetic features or single combinations of epigenetic features of interest the epigenetic feature selection criterion utilised by the computer code may be the average degree of methylation of the given epigenetic feature or set of epigenetic features on a given subset of samples. In one preferred embodiment, the epigenetic feature selection criterion utilised by the computer code may be the computation of the average degree of methylation on peripheral blood samples.

[0047] In preferred embodiments in which the candidate set of epigenetic features of interest comprises single combinations of epigenetic features of interest the epigenetic feature selection criterion may be the average pairwise correlation between the single epigenetic features.

[0048] In some preferred embodiments, the epigenetic features of interest and/or combinations of epigenetic features of interest selected by the computer code may be a defined number of the highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest. In other preferred embodiments the computer code selects all except a defined number of lowest ranking epigenetic features of interest and/or combinations of epigenetic features of interest. In yet other preferred embodiments, the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected or all except the epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score lesser than a defined threshold are selected by the computer code.

[0049] In preferred embodiments, the computer code repeats the feature selection steps iteratively until a defined number of epigenetic features of interest and/or combinations of epigenetic features of interest are selected or until all epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected.

[0050] In preferred embodiments the computer code calculates the optimal number of epigenetic features of interest and/or combinations of epigenetic features of interest and/or the optimal feature selection criterion score threshold by crossvalidation of a machine learning classifier on test subsets of the epigenetic feature data.

[0051] In some embodiments of the invention, the computer code uses the feature data set

corresponding to the defined new set of epigenetic features of interest to train a machine learning classifier algorithm.

BRIEF DESCRIPTION OF THE DRAWINGS

[0052] Figure 1 illustrates one embodiment of a process for epigenetic feature selection.

[0053] Figure 2 illustrates one embodiment of an iterative process for epigenetic feature selection.

[0054] Figure 3 shows the results of principal component analysis applied to methylation analysis data. The whole data set (25 samples) was projected onto its first 2 principal components. Circles represent cell lines, triangles primary patient tissue. Filled circles or triangles are AML, empty ones ALL samples.

[0055] Figure 4 Dimension dependence of feature selection performance. The plot shows the generalization performance of a linear SVM with four different feature selection methods against the number of selected features. The x-axis is scaled logarithmically and gives the number of input features for the SVM, starting with two. The y-axis gives the achieved generalization performance. Note that the maximum number of principle components corresponds to the number of available samples. Circles show the results for the Fisher Criterion, rectangles for t-test, diamonds for Backward Elimination and Triangles for PCA.

[0056] Figure 5 Fisher Criterion. The methylation profiles of the 20 highest ranking CpG sites according to the Fisher criterion are shown. The highest ranking features are on the bottom of the plot. The labels at the y -axis are identifiers for the CpG dinucleotide analyzed. The labels on the x - axis specify the phenotypic classes of the samples. High methylation corresponds to black, uncertainty to gray and low methylation to white.

[0057] Figure 6 Two sample t-test. The methylation profiles of the 20 highest ranking CpG sites according to the two sample t-test are shown. The highest ranking features are on the bottom of the plot. The labels at the y - axis are identifiers for the CpG dinucleotide analyzed. The labels on the x - axis specify the phenotypic classes of the samples. High methyla-

tion corresponds to black, uncertainty to gray and low methylation to white.

[0058] Figure 7 Backward elimination. The methylation profiles of the 20 highest ranking CpG sites according to the weights of the linear discriminant of a linear SVM are shown. The highest ranking features are on the bottom of the plot. The labels at the y - axis are identifiers for the CpG dinucleotide analyzed. The labels on the x - axis specify the phenotypic classes of the samples. High methylation corresponds to black, uncertainty to gray and low methylation to white.

[0059] Figure 8 Support Vector Machine on two best features of the Fisher criterion. The plot shows a SVM trained on the two highest ranking CpG sites according to the Fisher criterion with all ALL and AML samples used as training data. The black points are AML, the gray ones ALL samples. Circled points are the support vectors defining the white borderline between the areas of AML and ALL prediction. The gray value of the background corresponds to the prediction strength.

[0060] Figure 9 Likelihood ratio test for logistic regression models. The methylation profiles of the 12 highest ranking genomics regions according to the two sample likelihood ratio test are shown. The highest ranking features are on the bottom of the plot. Samples in categories A, B and C (normal colon tissue, colon tissue with inflammatory disease and colon polyps, respectively)were compared to samples of category D (colon cancer). Black indicates total methylation at a given CpG position, white represents no methylation at the particular position, with degrees of methylation represented in grey, from light (low proportion of methylation) to dark (high proportion of methylation).The figures to the right of the matrix show the p-values for comparison at each feature.

[0061] Figure 10 Average correlation scoring. The methylation profiles of the 12 highest ranking genomics regions according to the average between CpG correlation are shown. The highest ranking features are on the bottom of the plot. The degree of methylation at each position is shown by the shade of each position of the matrix, wherein black corresponds to high methylation and white corresponds to low methylation. Samples in categories A, B, C, D, E and F were compared to samples in category G. A, B, C, D, E, F and G are normal colon tissue, colon tissue with inflammatory disease, cancer samples from non-colon tissues, peripheral blood, normal tissues originating from non-colon sources, colon polyps and colon

cancer tissues respectively. The figures on the right side of the matrix show the correlation coefficient between the two groups.

DETAILED DESCRIPTION

[0062] The present invention provides methods and computer program products suitable for selecting epigenetic features comprising the steps of:

a) collecting and storing biological samples containing genomic DNA;
b) collecting and storing available phenotypic information about said biological samples; thereby defining a phenotypic data set;
c) defining at least one phenotypic parameter of interest;
d) using said defined phenotypic parameters of interest to divide said biological samples in at least two disjunct phenotypic classes of interest;
e) defining an initial set of epigenetic features of interest;
f) measuring and/or analyzing said defined epigenetic features of interest of said biological samples; thereby generating an epigenetic feature data set;
g) selecting those epigenetic features of interest and/or combinations of epigenetic features of interest that are relevant for epigenetically based prediction of said phenotypic classes of interest;
h) defining a new set of epigenetic features of interest based on the relevant epigenetic features of interest and/or combinations of epigenetic features of interest generated in step (g).

[0063] In the context of the present invention, "epigenetic features" are, in particular, cytosine methylations and further chemical modifications of DNA and sequences further required for their regulation. Further epigenetic parameters include, for example, the acetylation of histones which, however, cannot be directly analysed using the described method but which, in turn, correlates with DNA methylation. For illustration purpose the invention will be described using exemplary embodiments that analyze cytosine methylation.

Microarray-based DNA methylation analysis

[0064] In the first step of the method the genomic DNA must be isolated from the collected and stored biological samples. The biological samples may comprise cells, cellular compo-

nents which contain DNA or free DNA. Such sources of DNA may include cell lines, biopsies, blood, sputum, stool, urine, cerebral-spinal fluid, tissue embedded in paraffin such as tissue from eyes, intestine, kidney, brain, heart, prostate, lung, breast or liver, histologic object slides, and all possible combinations thereof. Extraction may be done by means that are standard to one skilled in the art, these include the use of detergent lysates, sonification and vortexing with glass beads. Such standard methods are found in textbook references (see, *e.g.*, Fritsch and Maniatis eds., Molecular Cloning: A Laboratory Manual, 1989. Once the nucleic acids have been extracted the genomic double stranded DNA is used in the analysis

[0065] Next, available phenotypic information about said biological samples is collected and stored. The phenotypic information may comprise, for example, kind of tissue, drug resistance, toxicology, organ type, age, life style, disease history, signaling chains, protein synthesis, behavior, drug abuse, patient history, cellular parameters, treatment history and gene expression. The phenotypic information for each collected sample will be preferably stored in a database.

[0066] At least one phenotypic parameter of interest is defined and used to divide the biological samples in at least two disjunct phenotypic classes of interest. For example the biological samples may be classified as ill and healthy, or tumor cell samples may be classified according to their tumor type or staging of the tumor type.

[0067] An initial set of epigenetic features of interest is defined. This initial set of epigenetic features of interest may be defined using preliminary knowledge data about their correlation with phenotypic parameters. In the illustrated preferred embodiments these epigenetic features of interest will be the cytosine methylation status at CpG dinucleotides located in the promoters, intronic and coding sequences of genes that are known to affect the chosen phenotypic parameters.

[0068] In the next step the cytosine methylation status of the selected CpG dinucleotides is measured. The state of the art method for large scale methylation analysis is described in PCT Application WO 99/28498. This method is based upon the specific reaction of bisulfite with cytosine which, upon subsequent alkaline hydrolysis, is converted to uracil which corresponds to thymidine in its base pairing behavior. However, 5-methylcytosine remains unmodified under these conditions. Consequently, the original DNA is converted in such a manner that

methylcytosine, which originally could not be distinguished from cytosine by its hybridization behavior, can now be detected as the only remaining cytosine using "normal" molecular biological techniques, for example, by amplification and hybridization to oligonucleotide arrays and sequencing. Therefore, in a preferred embodiment, DNA fragments of the pretreated DNA of regions of interest from promoters, intronic or coding sequence of the selected genes are amplified using fluorescently labeled primers. PCR primers can be designed complementary to DNA segments containing no CpG dinucleotides, thus allowing the unbiased amplification of methylated and unmethylated alleles. Subsequently the amplificates can be hybridized to glass slides carrying for each CpG position of interest a pair of immobilized oligonucleotides. These detection nucleotides are designed to hybridize to the bisulphite converted sequence around one CpG site which is either originally methylated (CG after pretreatment) or unmethylated (TG after pretreatment). Hybridization conditions have to be chosen to allow the detection of the single nucleotide differences between the TG and CG variants. Subsequently ratios for the two fluorescence signals for the TG and CG variants can be measured using, e.g., confocal microscopy. These ratios correspond to the degrees of methylation at each of the CpG sites tested.

[0069] Following these steps an epigenetic feature data set X has been generated containing the methylation status of all analyzed CpG dinucleotides. This data set may be represented as follows:

$$X = \{x^1, x^2, ..., x^i, ..., x^m\}, \text{ with}$$

$$x^i = \begin{bmatrix} x_1^i \\ x_2^i \\ \cdots \\ x_n^i \end{bmatrix},$$

wherein X is the methylation pattern data set for m samples,

x^i is the methylation pattern of sample i,

x_1^i to x_n^i are the CG/TG ratios for n analyzed CpG positions of sample j. x_1 to x_n de-

note the CG/TG ratios of the n CpG positions, the epigenetic features of interest.

Methylation based class prediction

[0070] The next step in large scale methylation analysis is to reveal by means of an evaluation algorithm the correlation of the methylation pattern with phenotypic classes of interest. The analysis strategy generally looks as follows. From many different DNA samples of known phenotypic class of interest (for example, from antibody-labeled cells of the same phenotype, isolated by immunofluorescence), methylation pattern data is generated in a large number of tests, and their reproducibility is tested. Then a machine learning classifier can be trained on the methylation data and the information which class the sample belongs to. The machine learning classifier can then with a sufficient number of training data learn, so to speak, which methylation pattern belongs to which phenotypic class. After the training phase, the machine learning classifier can then be applied to methylation data of samples with unknown phenotypic characteristic to predict the phenotypic class of interest this sample belongs to. For example, by measuring methylation patterns associated with two kinds of tissue, tumor or non-tumor, one obtains labeled data sets that can be used to build diagnostic identifiers.

[0071] In a preferred embodiment, where the samples are divided in two phenotypic classes of interest, the task of the machine learning classifier would be to learn, based on the methylation pattern for a given set of training examples $X = \{x^i : x^i \in R^n\}$ with known class membership $Y = \{y^i : y^i \in \{a,b\}\}$, where n is the number of CpGs, a and b are the two classes of interest, a discriminant function $f: R^n \rightarrow \{a,b\}$. This discriminant function can then be used to predict the classification of another data set $\{X'\}$ In machine learning nomenclature the percentage of miss-classifications of f on the training set $\{X,Y\}$ is called training error and is usually minimized by the learning machine during the training phase. However, what is of practical interest is the capability to predict the class of previously unseen samples, the so called generalization performance of the learning machine. This performance is usually estimated by the test error, which is the percentage of misclassifications on an independent test set $\{X'',Y''\}$ with known classification. The expected value of the test error for all independent test sets is called the risk.

[0072] The major problem of training a learning machine with good generalization perform-

ance is to find a discriminant function *f* which on the one hand is complex enough to capture the essential properties of the data distribution, but which on the other hand avoids over-fitting the data. Numerous machine learning algorithms, e.g., Parzen windows, Fisher's linear discrimant, two decision tree learners, or support vector machines are well known to those of skill in the art. The support vector machine (SVM) (Vapnik, V., Statistical Learning Theory, Wiley, New York, 1998) is a machine learning algorithm that has shown outstanding performance in several areas of application and has already been successfully used to classify mRNA expression data (see, *e.g.*, Brown, M., *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. USA, 97, 262-267, 2000). Therefore, in a preferred embodiment a support vector machine will be trained on the methylation data.

Feature selection

[0073] The major problem of all classification algorithms for methylation analysis is the high dimension of the input space, i.e. the number of CpGs, compared to the small number of analyzed samples. The classification algorithms have to cope with very few observations on very many epigenetic features. Therefore, the performance of classification algorithms applied directly to large scale methylation analysis data is generally poor.

[0074] The present invention provides methods and computer program products to reduce the high dimension of the methylation data by selecting those epigenetic features or combinations of epigenetic features that are relevant for epigenetically based classification. In this context, an epigenetic feature or a combination of epigenetic features is called relevant, if the accuracy and/or the significance of the epigenetically based classification is likely to decrease by exclusion of the corresponding feature data. For a given classifier, accuracy is the probability of correct classification of a sample with unknown class membership, significance is the probability that a correct classification of a sample was not caused by chance.

[0075] Figure 1 illustrates a preferred process for the selection of epigenetic features, preferably in a computer system. Epigenetic feature data is inputted in the computer system (1). The epigenetic feature dataset is grouped in at least two disjunct classes of interest, e.g., healthy cell samples and cancer cell samples. If the epigenetic feature data is grouped in more than two disjunct classes of interest pairs of classes or unions of pairs of classes are selected and

the feature selection procedure is applied to each of these pairs (2), (3). The reason to look at pairs of classes is that most machine learning classifiers are binary classifiers. Next (4) candidate sets of epigenetic features of interest and/or combinations of epigenetic features of interest are defined. These candidate features are ranked according to a defined feature selection criterion (5) and the highest ranking features are selected (6).

[0076] Figure 2 illustrates an iterative process for the selection of epigenetic features. The process is also preferably performed in a computer system. Epigenetic feature data, grouped in at least two disjunct classes of interest is inputted in the computer system (1). Pairs of disjunct classes or pairs of unions of disjunct classes are selected (2) and (3). Candidate sets of epigenetic features of interest and/or combinations of epigenetic features of interest are defined (4). The candidate features are ranked according to a defined feature selection criterion (5) and the highest ranking features are selected (6). If the number of the selected features is still too big, steps (4), (5) and (6) are repeated starting with the epigenetic feature data corresponding to the selected features of interest selected in step (6). This procedure can be repeated until the desired number of epigenetic features is selected. In every iterative step different candidate feature subsets and different feature selection criteria can be chosen.

[0077] In the following preferred embodiments for defining candidate sets of epigenetic features of interest or combinations of epigenetic features of interest and for defining a feature selection criteria to rank these candidate features will be described in detail.

Candidate feature sets

[0078] The canonical way to select all relevant features of interest would be to evaluate the generalization performance of the learning machine on every possible feature subset. This could be done by choosing every possible feature subset for a given set of epigenetic features and estimating the generalization performance by cross-validation on the training dataset. However, what makes this exhaustive search of the feature space practically useless is the enormous number of $\sum_{k=0}^{n}\binom{n}{k} = 2^n$ different feature combinations. Therefore, in a preferred embodiment, the present invention applies a two step procedure for feature selection. First, from the given set of epigenetic features candidate subsets of epigenetic features of interest or combinations of epigenetic features of interest are defined and then ranked according to a

chosen feature selection criterion.

[0079] In a preferred embodiment, the candidate set of epigenetic features of interest is the set of all subsets of the given epigenetic feature set. In another preferred embodiment, the candidate set of epigenetic features of interest is the set of all subsets of a defined cardinality, i.e. the set of all subsets with a given number of elements. Particularly, the candidate set of epigenetic features of interest is chosen to be the set of all subsets of cardinality 1, i.e. every single feature is selected and ranked according to the defined feature selection criterion.

[0080] In other preferred embodiments, dimension reduction techniques are applied to define combinations of epigenetic features of interest. In a preferred embodiment, principal component analysis (PCA) is applied to the epigenetic feature data set. As known to one skilled in the art, for a given data set X, principal component analysis constructs a set of orthogonal vectors (principal components) which correspond to the directions of maximum variance in the data. The single linear combination of the given features that has the highest variance is the first principal component. The highest variance linear combination orthogonal to the first principal component is the second principal component, and so forth (see, *e.g.*, Mardia, K.V., *et. al,* Multivariate Analysis, Academic Press, London, 1979). To define the candidate set of combinations of epigenetic features of interest the first principal components are chosen.

[0081] In another preferred embodiment, multidimensional scaling (MDS) is used to define the candidate features. Contrary to PCA which finds a low dimensional embedding of the data points that best preserves their variance, MDS is a dimension reduction technique that finds an embedding that preserves the interpoint distances (see, *e.g.*, Mardia, K.V., *et al,* Multivariate Analysis, Academic Press, London, 1979). To define the candidate set of epigenetic features the epigenetic feature data set X is embedded with MDS in a d -dimensional vector space, the calculated coordinate vectors defining the candidate features. The dimension d of this space is can be fixed and supplied by a user. If not given, one way to estimate the true dimensionality d of the data is to vary d from 1 to n and calculate for every embedding the residual variance of the data. Plotting the residual variance versus the dimension of the embedding the curve generally decreases as the dimensionality d is increased but shows a characteristic "elbow" at which the curve ceases to decrease significantly with added dimensions. This point gives the true dimension of the data (see, *e.g.,* Kruskal, J.B., Wish, M., Multidimensional Scaling, Sage University Paper Series on Quantitative Applications in the Social

Sciences, London, 1978, Chapter 3). In another preferred embodiment isometric feature mapping is applied as dimensional reduction technique. Isometric feature mapping is a dimension reduction approach very similar to MDS in searching for a lower dimensional embedding of the data that preserves the interpoint distances. However, contrary to MDS isometric feature mapping can cope with nonlinear structure in the data. The isometric feature mapping algorithm is described in Tenenbaum, J. B., A Global Geometric Framework for Nonlinear Dimensionality reduction, Science 290, 2319-2323, 2000. For the definition of the candidate features, the epigenetic feature data set is embedded in d dimensions using the isometric feature mapping algorithm, the coordinate vectors in the d -dimensional space defining the candidate features. The dimensionality d of the embedding can be fixed and supplied by a user or an optimal dimension can be estimated by looking at the decrease of residual variance of the data for embeddings in increasing dimensions as described for MDS.

[0082] In another preferred embodiment, cluster analysis is used to define the candidate set of epigenetic features. Cluster analysis is an effective means to organize and explore relationships in data. Clustering algorithms are methods to divide a set of m observations into g groups so that members of the same group are more alike than members of different groups. If this is successful, the groups are called clusters. Two types of clustering, k-means clustering or partitioning methods and hierarchical clustering, are particularly useful for use with methods of the invention. In signal processing literature partitioning methods are generally denoted as vector quantisation methods. In the following we will use the term k-means clustering synonymously with partitioning methods and vector quantisation methods. k-means clustering partitions the data into a preassigned number of k groups. k is generally fixed and provided by a user. An object (such as a the methylation pattern of a sample) can only belong to one cluster. k-means clustering has the advantage that points are re-evaluated and errors do not propagate. The disadvantages include the need to know the number of clusters in advance, assumption that clusters are round and assumption that the clusters are the same size. Hierarchical clustering algorithms have the advantage to avoid specifying how many clusters are appropriate. They provide the user with many different partitions organized as a tree. By cutting the tree at some level the user may choose an appropriate partitioning. Hierarchical clustering algorithms can be divided in two groups. For a set of m samples, agglomerative algorithms start with m clusters. The algorithm then picks the two clusters with the smallest dissimilarity and merges them. This way the algorithm constructs the tree so to speak from the bottom up.

Divisive algorithms start with one cluster and successively split clusters into two parts until this is no longer possible. These algorithms have the advantage that if most interest is on the upper levels of the cluster tree they are much more likely to produce rational clusterings their disadvantage is very low speed. Compared to k-means clustering hierarchical clustering algorithms suffer from early error propagation and no re-evaluation of the cluster members. A detailed description of clustering algorithms can be found in, *e.g.,* Hartigan, J.A., Clustering Algorithms, Wiley, New York, 1975). Having subjected the epigenetic feature data set X to a cluster analysis algorithm, all epigenetic features belonging to the same cluster are combined, e.g., the cluster mean is chosen to represent all features belonging to the same cluster, to define the candidate features.

[0083] A preferred means of practising the invention is to define the candidate set of epigenetic features according to a priori biological information and group epigenetic features of interest with similar biological properties together. Epigenetic features may be combined or grouped according to any biological properties that enable an assumption that the members of the candidate set have similar or correlated epigenetic status , most preferably methylation status known in the art as 'co-methylation'. Wherein this is not known it may be inferred using any parameters which may be used as reasonable indicators that members of a set of CpG positions have a common methylation status, which may in particular be selected from the group consisting of:

- Proximity to each other; wherein the epigenetic features are close enough that it may be assumed or expected that they have similar or correlated epigenetic status. In particular, when the epigenetic features belong to the same CpG island (defined as a sequence greater than 200 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65 or greater); and

- Associated function: epigenetic features belonging to genes that are known to have similar function and/or co-regulated and/or belong to the same biological pathway and/or have sequence similarity and therefore expected to be regulated by similar transcription factors. This is particularly advantageous because then epigenetic features that have a strong covariance structure are analyzed together which increases the, power i.e. the probability of identifying a relevant epigenetic feature. This can also be advantageous when certain technical properties of feature combinations are preferred for further assay development.

[0084] It has to be stressed that in the present invention the described statistical analysis methods aren't used for a final analysis of the large scale methylation data. They are used to define candidate sets of relevant epigenetic features of interest which are then further analyzed to select the relevant epigenetic features. These relevant epigenetic features of interest are than used in subsequent analysis.

Feature selection criteria

[0085] Having defined a candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest, the candidate features are ranked according to preferred selection criteria. In the machine learning literature the feature selection methods are generally distinguished in *wrapper* methods and *filter* methods. The essential difference between these approaches is that a wrapper method makes use of the algorithm that will be used to build the final classifier, while a filter method does not. A filter method attempts to rank subsets of the features by making use of sample statistics computed from the empirical distribution.

[0086] Some embodiments of the invention make use of wrapper methods. In a preferred embodiment the feature selection criterion may be the training error of a machine learning classifier trained on the epigenetic feature data corresponding to the chosen candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest. For example, if the candidate set of epigenetic features of interest was chosen to be the set of all two-CpG-combinations of the n given CpG positions analyzed, i.e.,

$$\{\{x_1, x_2\}, \{x_1, x_3\}, \ldots, \{x_1, x_n\}, \ldots, \{x_{n-1}, x_n\}\}$$

a machine learning classifier is trained for every of the $\binom{n}{2}$ two-CpG-combinations on the corresponding methylation pattern data $X = \{x^i : x^i \in R^2\}$ with known class membership $Y = \{y^i : y^i \in \{a, b\}\}$, and the percentage of misclassifications determined. The two-CpG-subsets are ranked with increasing error.

[0087] In another preferred embodiment the feature selection criterion may be the risk of the machine learning classifier trained on the epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of

interest. The risk is the expected test error of a trained classifier on independent test sets $\{X', Y'\}$. As known to one skilled in the art a common method to determine the test error of a classifier is cross-validation (see, *e.g.,* Bishop, C., Neural networks for pattern recognition, Oxford University Press, New York, 1995). For cross-validation the training set $\{X, Y\}$ is divided into several parts and in turn using one part as test set, the other parts as training sets. A special form is leave-one-out cross-validation where in turn one sample is dropped from the training set and used as test sample for the classifier trained on the remaining samples. Having evaluated the risk by cross-validation for every element of the defined candidate set of epigenetic features and/or combinations of epigenetic features the elements are ranked by increasing risk.

[0088] If for the applied machine learning classifier theoretical bounds on the risk can be given, these bounds can be chosen as feature selection criteria. A preferred classifier for the analysis of methylation data is the support vector machine algorithm (SVM). For the SVM algorithm bounds on the risk can be derived from statistical learning theory. Details can be found in Vapnik, V. Statistical Learning Theory, Wiley, New York, 1998 or Cristianini, N., Shaw-Taylor, J., An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000. For example, a bound (Theorem 4.24 in Cristianini, Shaw-Taylor) that can be applied as feature selection criterion states that with probability 1-d the risk r of the SVM classifier is bound by

$$r \leq \frac{c}{l}\left(\frac{R^2 + z^2 \log(1/D)}{D^2}\right)\log^2(l) + \log\left(\frac{1}{l}\right)$$

wherein c is a constant, l is the number of training samples, R is the radius of the minimal sphere enclosing all data points, D is the margin of the support vectors and z is the margin slack vector. R, D, and z are easily derived when training the SVM on every candidate feature subset. Therefore the candidate feature subsets can be ranked with increasing bound values.

[0089] Other preferred embodiments of the invention make use of filter methods. If the candidate set of epigenetic features as defined in the preliminary step of the feature selection method of the invention is a set consisting of single epigenetic features combinations of epigenetic features, i.e. $\{\{z_1\}\{z_2\}\{z_3\}...\}$ where the z_i are epigenetic features x_i or combinations of

single epigenetic features, test statistics computed from the empirical distribution can be chosen as epigenetic feature selection criteria. A preferred test statistic is a t-test. For example, if the analyzed samples can be divided in two classes, say ill and healthy, for every single CpG position x_i, the null hypothesis, that the methylation status class means are the same in both classes can be tested with a two sample t-test. The CpG positions can than be ranked by increasing significance value. If there are doubts that the methylation status distribution for any CpG can be approximated by a Gaussian normal distribution other embodiments are preferred that use rank test, particularly a Wilcoxon rank test (see, *e.g.,* Mendenhall, W, Sincich, T, Statistics for engineering and the sciences, Prentice-Hall, New Jersey, 1995).

[0090] In another preferred embodiment the significance value of a multivariate statistical test is applied to a subgroup/combination of a several epigenetic features. Any suitable test may be applied, however a preferred test statistic is the T^2-test, and other similar test statistics such as, but not limited to the likelihood ratio test for logistic regression models.

[0091] In another preferred embodiment, the Fisher criterion is chosen as feature selection criterion.

[0092] The Fisher criterion is a classical measure to assess the degree of separation between two classes (see, e.g., Bishop, C., Neural networks for pattern recognition, Oxford University Press, New York, 1995). If, for example, the samples can be divided in two classes, say A and B, the discriminative power of the *k*th CpG x_k is given as:

$$J(k)=\frac{\left(m_k^A - m_k^B\right)}{\left(s_k^{2A} + s_k^{2B}\right)},$$

wherein $m_k^{A/B}$ is the mean and $s_k^{A/B}$ is the standard deviation of all sample data values x_k^i with $Y^j = A/B$. The Fisher criterion gives a high ranking for CpGs where the two classes are far apart compared to the within class variances.

[0093] In another preferred embodiment the weights of a linear discriminant used as the classifier are used as the feature selection criterion. The concept of linear discriminant functions is well known to one skilled in the art of neural network and pattern recognition. A detailed introduction can be found, for example, in Bishop, C., Neural networks for pattern recognition,

Oxford University Press, New York, 1995. In short, for a two-category classification, if x^j is the methylation pattern of sample j, a linear discriminant function $z : R^n \rightarrow R$ has the form:

$$z(x^j)=w^T x^j + w_0.$$

[0094] The pattern x^j is assigned to class C_1 if $z(x^j)>0$ and to class C_2 if $z(x^j)\leq 0$. The n-dimensional vector w is called the *weight vector* and the parameter w_0 the *bias*. To estimate the weight vector, the discriminant function is trained on a training set. The estimation of the weight vector may, for example, be done calculating a least-squares fit on a training set. Having estimated the coordinate values of the weight vectors, the features can be ranked according to the size of the weight vector coordinates. In a preferred embodiment the weight vector is estimated by Fisher's linear discriminant:

$$w \propto S_W^{-1}(m_2 - m_1)$$

where m_1 and m_2 are the mean vectors of the two classes

$$m_1 = \frac{1}{N_1}\sum_{i\in C_1} x^i,\ m_2 = \frac{1}{N_2}\sum_{i\in C_2} x^i$$

and

$$S_W = \sum_{i\in C_1}(x^i - m_1)(x^i - m_1)^T + {}^T + \sum_{i\in C_2}(x^i - m_2)(x^i - m_2)^T$$

is the total *within-class* covaiance matrix.

[0095] Another preferred embodiment uses the support vector machine (SVM) algorithm to estimate the weight vector w, see Vapnik, V., Statistical Learning Theory, Wiley, New York, 1998, for a detailed description.

[0096] In another preferred embodiment PCA is used to rank the defined candidate epigenetic features in the following way: The epigenetic feature data corresponding to the defined candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest is subject to principal component analysis (PCA). Then the ranks of the weights of the first principal component are used to rank the candidate features.

[0097] In yet another preferred embodiment, the feature selection criterion is the mutual information between the phenotypical classes of the sample and the classification achieved by an optimally selected threshold on every candidate feature. If $\{\{z_1\}\{z_2\}\{z_3\}...\}$ is the defined set of candidate features where the z_i are single epigenetic features x_i or combinations of single epigenetic features x_i, for every z_i a simple classifier is defined by assigning sample j to class C_1 if $z_i^j < b_i$ and to class C_2 if $z_i^j \leq b_i$. The threshold b_i is chosen such as to maximize the number of correct classifications on the training data. Note that for every candidate feature the optimal threshold is determined separately. To rank the candidate features the mutual information between each of these classifications and the correct classification is calculated. As known to one skilled in the art the mutual information I of two random variables r and s is given by

$$I(r, s) = \mathrm{H}(r) + \mathrm{H}(s) - \mathrm{H}(r, s).$$

$$\mathrm{H}(r) = -\sum_i p_i \ln p_i$$

is the entropy of random variable taking the discrete values with probability and

$$\mathrm{H}(r, s) = -\sum_{ij} p_{ij} \ln p_{ij}$$

is the joint entropy of the random variables r and s taking the values r_i and s_j with probability p_{ij} (see, *e.g.*, Papoulis, A., Probability, Random Variables and Stochastic Processes, McGraw-Hill, Boston, 1991). In a preferred embodiment, this last step of calculating the mutual information is omitted and the candidate features are ranked according to the number of correct classifications their corresponding optimal threshold classifiers achieve on the training data.

[0098] Another preferred embodiment for the choice of the feature selection criterion can be used if the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest has been defined to be the principal components, subjecting the epigenetic feature data set to PCA as described in the previous section. Then these candidate features can be simply ranked according to the absolute value of the eigenvalues of the principal components.

Selecting the most important features

[0099] Having defined the candidate set of epigenetic features of interest and/or combinations of epigenetic features of interest and ranked theses candidate features according to a preferred feature selection criterion as described in the preceding sections, the final step of the method is to select the most important features from the candidate set.

[00100] In a preferred embodiment, a defined number k of highest ranking epigenetic features of interest and/or combinations of epigenetic features of interest is selected from the candidate set. k can be fixed and hard coded in the computer program product or supplied by a user. In another preferred embodiment, all except a defined number k of lowest ranking epigenetic features of interest and/or combinations of epigenetic features of interest are selected from the candidate set. k can be fixed and hard coded in the computer program product or supplied by a user.

[00101] In other preferred embodiments, all epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score greater than a defined threshold are selected. The threshold can be fixed and hard coded in the computer program. Or, preferred when using the filter methods, the threshold is calculated from a predefined quality requirement like a significance threshold using the empirical distribution of the data. Or, further preferred, the threshold value may be supplied by a user. In other preferred embodiments all epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection criterion score lesser than a defined threshold are selected, the threshold being fixed and hard coded in the computer program, calculated from the empirical distribution and predefined quality requirements or provided by a user.

[00102] In other preferred embodiments, the feature selection steps are iterated until a defined number of epigenetic features of interest and/or combinations of epigenetic features of interest are selected or until all epigenetic features of interest and/or combinations of epigenetic features of interest with a feature selection score greater than a defined threshold are selected. In every iterative step the same or another feature selection criterion could be chosen. In a similar manner the definition of the new candidate set to rank with the feature selection criterion can be the same in every iterative step or changing with the iterative steps.

[00103] A special form of an iterative strategy is known as *backward elimination* to one skilled in the art. Starting with the full set of epigenetic features as candidate feature set, the preferred feature selection criterion is evaluated and all features selected except the one with the smallest score. These steps are iteratively repeated with the new reduced feature set as candidate set until all except a defined number of features are deleted from the set or all feature with feature selection score lesser than a defined threshold are deleted. Another preferred iterative strategy is known as *forward selection* to one skilled in the art. Starting with the candidate feature set of all single features, for example, $\{\{x_1\}\{x_2\}\{x_3\}..\{x_n\}\}$ the single features are ranked according to the chosen features selection criterion and all are selected for the next iterative step. In the next step the candidate set chosen is the set of subsets of cardinality 2 that include the highest ranking feature from the preceding step. Suppose $\{x_3\}$ is the highest ranking single feature, the candidate set of features of interest will be chosen as $\{\{x_3,x_1\}\{x_3,x_2\}\{x_3,x_4\}..\{x_3,x_n\}\}$. The feature selection criterion is evaluated and the subset that gives the largest increase in score forms the basis of the candidate set of subsets of cardinality 3 defined in the next iterative step. These steps are repeated until a fixed or user defined cardinality is reached or until there is no further increase in feature selection criterion score from one step to the next.

[00104] Another preferred embodiment uses a machine learning classifier to determine the optimal number of epigenetic features of interest and/or combinations of epigenetic features of interest to select. The test error of the classifier is evaluated by cross-validation using in the first stage only the data for the highest ranking feature or feature combination and adding in each successive step one additional feature or feature combination according to the ranking.

[00105] Having used the methods of the invention for epigenetic feature selection, the epigenetic feature data corresponding to the selected epigenetic features or combinations of epigenetic features can be used to train a machine learning classifier for the given classification problem. New data to be classified by the trained machine would be pre-processed with the same feature selection method as the training set, before inputting to the classifier. As the example in the following section shows, the methods of the invention greatly improve the performance of machine learning classifiers applied to large scale methylation analysis data.

Example 1

[00106] This example illustrates some embodiments of the method of the invention and its application in DNA methylation based cancer classification. Samples obtained from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) and cell lines derived from different subtypes of leukemias were chosen to test if classification can be achieved solely based on DNA methylation patterns.

Experimental protocol

[00107] High molecular chromosomal DNA of 6 human B cell precursor leukaemia cell lines, 380, ACC 39; BV-173, ACC 20; MHH-Call-2, ACC 341; MHH-Call-4, ACC 337; NALM-6, ACC 128; and REH, ACC 22 were obtained from the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig). DNA prepared from 5 human acute myeloid leukaemia cell lines CTV-1, HL-60, Kasumi-1, K-562 (human chronic myeloid leukaemia in blast crisis) and NB4 (human acute promyelocytic leukaemia) were obtained from University Hospital Charite, Berlin. T cells and B cells from peripheral blood of 8 healthy individuals were isolated by magnetically activated cell separation system (MACS, Miltenyi, Bergisch-Gladbach, Germany) following the manufacturer's recommendations. As determined by FACS analysis, the purified CD4+ T cells were >73 % and the CD19+ B cells > 90 %. Chromosomal DNA of the purified cells was isolated using QIAamp DNA minikit (Qiagen, Hilden, Germany) according to the recommendation of the manufacturer. DNA isolated at time of diagnosis of the peripheral blood or bone marrow samples of 5 ALL-patients (acute lymphoid leukaemia) and 3 AML-patients (acute myeloid leukaemia) was obtained from University Hospital Charite, Berlin.

[00108] 81 CpG dinucleotide positions located in CpG rich regions of the promoters, intronic and coding sequences of the 11 genes ELK1, CSNK2B, MYCL1, CD63, CDC25A, TUBB2, CD1A, CDK4, MYCN, AR and c-MOS were chosen to be analyzed. The 11 genes were randomly selected from a panel of genes representing different pathways associated with tumorigenesis. Total DNA of all samples was treated using a bisulfite solution as described in A. Olek, J. Oswald, J. Walter, Nucleic Acid Res. 24, 5064 (1996). The genomic DNA was digested with MssI (MBI Fermentas, St. Leon-Rot, Germany) prior to the modification by bisulphite. For the PCR amplification of the bisulphite treated sense strand of the 11 genes

primers were designed according to the guidelines of Clark and Frommer (S. J. Clark, M. Frommer, in Laboratory Methods for the Detection of Mutations and Polymorphisms in DNA, G. R. Taylor ed., CRC Press, Boca Raton 1997). The PCR primers were designed complementary to DNA segments containing no CpG dinucleotides. This allowed unbiased amplification of both methylated and unmethylated alleles in one reaction. 10 ng DNA was used as template DNA for the PCR reactions. The template DNA, 12.5 pmol or 40 pmol (CY5-labelled) of each primer, 0.5-2 U Taq polymerase (HotStarTaq, Qiagen, Hilden, Germany) and 1 mM dNTPs were incubated with the reaction buffer supplied with the enzyme in a total volume of 20 µl. After activation of the enzyme (15 min, 96 °C) the incubation times and temperatures were 95°C for 1 min followed by 34 cycles (95°C for 1 min, annealing temperature (see Supplementary information) for 45 sec, 72°C for 75 sec) and 72°C for 10 min.

[00109] Oligonucleotides with a C6-amino modification at the 5′end were spotted with 4-fold redundancy on activated glass slides (T. R. Golub et al., Science 286, 531, 1999). For each analyzed CpG position two oligonucleotides N(2-16)-CG-N(2-16) and N(2-16)-TG-N(2-16), reflecting the methylated and non methylated status of the CpG dinucleotides, were spotted and immobilized on the glass array. The oligonucleotide microarrays representing 81 CpG sites were hybridized with a combination of up to 11 Cy5-labelled PCR fragments as described in D. Chen, Z. Yan, D. L. Cole, G. S. Srivatsa, Nucleic Acid Res 27, 389, 1999. Hybridization conditions were selected to allow the detection of the single nucleotide differences between the TG and CG variants. Subsequently, the fluorescent images of the hybridized slides were obtained using a GenePix 4000 microarray scanner (Axon Instruments). Hybridization experiments were repeated at least three times for each sample.

[00110] Average log CG/TG ratios of the fluorescent signals for the 81 CpG positions were calculated.

Methylation based class prediction

[00111] Next support vector machines were trained on this methylation data to learn the classification of samples obtained from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML).

[00112] In order to evaluate the prediction performance of these SVMs a cross-

validation method (Bishop, C., Neural networks for pattern recognition, Oxford University Press, New York, 1995) was used. For each classification task, the 25 samples were partitioned into 8 groups of approximately equal size. Then the SVM predicted the class for the test samples in one group after it had been trained using the 7 other groups. The number of misclassifications was counted over 8 runs of the SVM algorithm for all possible choices of the test group. To obtain a reliable estimate for the test error the number of misclassifications were averaged over 50 different partitionings of the samples into 8 groups.

[00113] First, two SVM were trained using all 81 CpG positions as separate dimension. As can be seen in Table I the SVM with linear kernel trained on this 81 dimensional input space had an average test error of 16%. Using a quadratic kernel did not significantly improve the results. An obvious explanation for this relatively poor performance is that we have only 25 data points (even less in the training set) in a 81 dimensional space. Finding a separating hyperplane under these conditions is a heavily under-determined problem. This shows the poor performance of machine learning classifiers applied to large scale methylation analysis data and the great need for the methods provided by the described invention.

Epigenetic feature selection

[00114] Subsequently some of the preferred embodiments of the invention for selecting epigenetic features were applied and the performance of the SVM for this reduced feature set tested using cross-validation as described above.

[00115] First, PCA was used for epigenetic feature selection. The methylation data for all 81 CpG positions was subject to PCA and the first k principle components selected for $k = 2$ and $k = 5$. Table I shows the results of the performance of SVMs trained and tested on the methylation data projected on this 2- and 5-dimensional feature space. For $k = 2$ the SVM with linear kernel had an average test error of 21% for $k = 5$ an average test error of 28%. The results for a SVM with quadratic kernel were even worse. The reason for this poor performance is that PCA does not necessarily extract features that are important for the discrimination between ALL and AML. It first picks the features with the largest variance, which are in this case discriminating between cell lines and primary patient tissue (see Figure 3), i.e. subgroups that are not relevant to the classification. As shown in Figure 4 features carrying information about the leukemia subclasses appear only from the 9^{th} principal component on.

[00116] Next, all 81 CpG positions were ranked using the Fisher criterion to determine the discriminative power of each CpG for the classification of ALL versus AML. Figure 5 shows the methylation profiles of the best 20 CpGs . The score increases from bottom to top. SVMs were trained on the 2 and 5 highest ranking CpGs The test error is shown in Table I. The results show a dramatic improvement of generalization performance compared to no feature selection or PCA. For 5 CpGs the test error decreases from 16% for the linear kernel SVM without feature selection to 3%. Figure 4 shows the dependence of generalization performance from the selected dimension k and indicates that especially Fisher criterion (circles) gives dimension independent good generalization for reasonable small k.

[00117] The highest ranking CpG sites according to a two sample t-test are shown in Figure 6. The ranking of the CpG is very similar to the Fisher criterion. The test errors for SVMs trained on the k highest ranking features for $k = 2$ and $k = 5$ are shown in Table I. Compared to the Fisher criterion the generalization performance is considerably worse.

[00118] Furthermore the weights of the linear discriminant of the support vector machine algorithm were chosen as feature selection criterion. The candidate features were defined using the *backward elimination* strategy. The SVM with linear kernel was trained on all 81 CpG and the normal vector of the separating hyperplane the SVM uses for discrimination calculated. The feature ranking is then simply given by the absolute value of the components of the normal vector. The feature with the smallest component was deleted and the SVM retrained on the reduced feature set. This procedure is repeated until the feature set is empty. The methylation pattern for the highest ranking CpGs according to this selection method is shown in Figure 7. The ranking differs considerably from the Fisher ant t-test rankings. However, as shown in Table I the generalization results evaluated when training the SVM on the 2 or 5 highest ranking features weren't better than for the Fisher criterion although this method is computationally much more expensive than calculating the Fisher criterion.

[00119] Finally, the space of all two feature combinations was exhaustively searched to find the optimal two features for classification by evaluating the generalization performance of the SVM using cross-validation. For every of the $\binom{81}{2} = 3240$ two CpG combination the leave-one out cross-validation error of a SVM with quadratic kernel was calculated on the training set. From all CpG pairs with minimum leave-one-out error the one with the smallest

radius margin ratio was selected. This pair was considered to be the optimal feature combination and was used to evaluate the generalization performance of the SVM on the test set. The average test error of the exhaustive search method was with 6% the same as the one of the Fisher criterion in the case of two features and a quadratic kernel. For five features the exhaustive computation is already infeasible. In the absolute majority of cross-validation runs the CpGs selected by exhaustive search and Fisher criterion were identical. In some cases suboptimal CpGs were chosen by the exhaustive search method.

[00120] It follows that at least for this data set the simple Fisher criterion is the preferable technique for epigenetic feature selection.

[00121] This example clearly shows that microarray based methylation analysis combined with supervised learning techniques and the methods of this invention can reliably predict known tumor classes. Figure 8 shows the result of the SVM classification trained on the two highest ranking CpG sites according to the Fisher criterion.

Example 2

[00122] In the following samples obtained from patients with colon cancer were chosen to test if classification can be achieved solely based on DNA methylation patterns.

[00123] DNA samples were extracted using lysis buffer from Qiagen and the Roche magnetic separation kit for genomic DNA isolation. DNA samples were also extracted using Qiagen Genomic Tip-100 columns, as well as the MagnaPure device and Roche reagents. All samples were quantitated using spectrophotometric or fluorometric techniques and on agarose gels for a subset of samples.

Bisulfite treatment and mPCR

[00124] Total genomic DNA of all samples was bisulfite treated converting unmethylated cytosines to uracil. Methylated cytosines remained conserved. Bisulfite treatment was performed with minor modifications according to the protocol described in Olek et al. (1996). In order to avoid processing all samples with the same biological background together resulting in a potential process-bias in the data later on, the samples were randomly grouped into processing batches. For bisulfite treatment we created batches of 50 samples randomized for

sex, diagnosis, and tissue. Per DNA sample two independent bisulfite reactions were performed. After bisulfitation 10 ng of each DNA sample was used in subsequent mPCR reactions containing 6-8 primer pairs.

a. Each reaction contained the following:
b. 0.4 mM each dNTPS
c. 1 Unit Taq Polymerase
d. 2.5 µl PCR buffer
e. 3.5 mM MgCl2
f. 80 nM Primerset (12-16 primers)
g. 11.25 ng DNA (bisulfite treated)

Forty cycles were carried out as follows: Denaturation at 95°C for 15 min, followed by annealing at 55°C for 45 sec., primer elongation at 65°C for 2 min. A final elongation at 65°C was carried out for 10 min.

1.1.2 Hybridization

[00125] All PCR products from each individual sample were then hybridised to glass slides carrying a pair of immobilised oligonucleotides for each CpG position under analysis. Each of these detection oligonucleotides was designed to hybridise to the bisulphite converted sequence around one CpG site which was either originally unmethylated (TG) or methylated (CG). Hybridisation conditions were selected to allow the detection of the single nucleotide differences between the TG and CG variants.

[00126] 5 µl volume of each multiplex PCR product was diluted in 10 x Ssarc buffer (10 x Ssarc:230 ml 20 x SSC, 180 ml sodium lauroyl sarcosinate solution 20%, dilute to 1000 ml with dH2O). The reaction mixture was then hybridised to the detection oligonucleotides as follows. Denaturation at 95°C, cooling down to 10 °C, hybridisation at 42°C overnight followed by washing with 10 x Ssarc and dH2O at 42°C.

[00127] Fluorescent signals from each hybridised oligonucleotide were detected using genepix scanner and software. Ratios for the two signals (from the CG oligonucleotide and the TG oligonucleotide used to analyse each CpG position) were calculated based on comparison of intensity of the fluorescent signals.

[00128] The samples were processed in batches of 80 samples randomized for sex, diagnosis, tissue, and bisulphite batch For each bisulfite treated DNA sample 2 hybridizations were performed. This means that for each sample a total number of 4 chips were processed.

Data analysis methods

[00129] Analysis of the chip data: From raw hybridization intensities to methylation ratios; The log methylation ratio (log(CG/TG)) at each CpG position is determined according to a standardized preprocessing pipeline that includes the following steps:

[00130] For each spot the median background pixel intensity is subtracted from the median foreground pixel intensity (this gives a good estimate of background corrected hybridization intensities):

For both CG and TG detection oligonucleotides of each CpG position the background corrected median of the 4 redundant spot intensities is taken;

For each chip and each CpG position the log(CG/TG) ratio is calculated;

For each sample the median of log(CG/TG) intensities over the redundant chip repetitions is taken.

This ratio has the property that the hybridization noise has approximately constant variance over the full range of possible methylation rates (Huber et al., 2002).

Hypothesis testing

[00131] The main task is to identify markers that show significant differences in the average degree of methylation between two classes. A significant difference is detected when the null-hypothesis that the average methylation of the two classes is identical can be rejected with $p<0.05$. Because we apply this test to a whole set of potential markers we have to correct the p-values for multiple testing. This was done by applying the Bonferroni method.

[00132] For testing the null hypothesis that the methylation levels in the two classes are identical we used the likelihood ratio test for logistic regression models. The logistic regression model for a single marker is a linear combination of methylation measurements from all CpG positions in the respective genomic region of interest (ROI). A significant p-value for a

marker means that this ROI has some systematic correlation to the question of interest as given by the two classes.

Class prediction by supervised learning

[00133] In order to give a reliable estimate of how well the CpG ensemble of a selected marker can differentiate between different tissue classes we can determine its prediction accuracy by classification. For that purpose we calculate a methylation profile based prediction function using a certain set of tissue samples with their class label. This step is called training and it exploits the prior knowledge represented by the data labels. The prediction accuracy of that function is then tested by cross-validation or on a set of independent samples. As a method of choice, we use the support vector machine (SVM) algorithm to learn the prediction function. If not stated otherwise, for this report the risk associated with false positive or false negative classifications are set to be equal relative to the respective class sizes. It follows that the learning algorithm obtains a class prediction function with the objective to optimize accuracy on an independent test sample set. Therefore sensitivity and specificity of the resulting classifier can be expected to be approximately equal.

Estimating the performance of the tissue class prediction: Cross Validation

[00134] With limited sample size the cross-validation method provides an effective and reliable estimate for the prediction accuracy of a discriminator function and therefore in addition to the significance of the markers we provide cross-validation accuracy, sensitivity and specificity estimates. For each classification task, the samples were partitioned into 5 groups of approximately equal size. Then the learning algorithm was trained on 4 of these 5 sample groups. The predictor obtained by this method was then tested on the remaining group of independent test samples. The number of correct positive and negative classifications was counted over 5 runs for the learning algorithm for all possible choices of the independent test group without using any knowledge obtained from the previous runs. This procedure was repeated on up to 10 random permutations of the sample set. Note that the above-described cross-validation procedure evaluates accuracy, sensitivity and specificity using practically all possible combinations of training and independent test sets. It therefore gives a better estimate of the prediction performance than simply splitting the samples into one training sample set and one independent test set.

Results

[00135] Figure 9 shows an example for a multivariate ranking by a likelihood ratio test for logistic regression models. Every set of CpGs belonging to the same region of interest on the genome is assigned one p-value. Samples in categories A, B and C (normal colon tissue, colon tissue with inflammatory disease and colon polyps, respectively)were compared to samples of category D (colon cancer).

[00136] Figure 10 shows an example for a multivariate ranking by average between CpG correlation. Samples in categories A, B, C, D, E and F were compared to samples in category G. A, B, C, D, E, F and G are normal colon tissue, colon tissue with inflammatory disease, cancer samples from non-colon tissues, peripheral blood, normal tissues originating from non-colon sources, colon polyps and colon cancer tissues. Every set of CpGs belonging to the same region of interest on the genome is assigned one correlation coefficient. The higher this coefficient the higher the degree of co-methylation within this region.

Table I

	Training Error 2 Features	Training Error 2 Features	Training Error 5 Features	Training Error 5 Features
Linear Kernel				
Fisher Criterion	0,01	0,05	0,00	0,03
t-Test	0,05	0,13	0,00	0,08
Backward Estimation	0,02	0,17	0,00	0,05
PCA	0,13	0,21	0,05	0,28
No Feature Selection	0,00	0,16	-	-
Quadratic Kernel				
Fisher Criterion	0,00	0,06	0,00	0,03
t-Test	0,04	0,14	0,00	0,07
Backward Estimation	0,00	0,12	0,00	0,05
PCA	0,10	0,30	0,00	0,31
Exhaustive Search	0,00	0,06	-	-
No Feature Selection	0,00	0,15	-	-